A Study on Object Search and Relationship Search from Text Archive Data(Abstract_要旨)

AUTHOR(S):

Yating, Zhang

（ 続紙 １ ）

京都大学	博士（ 情報学 ）	氏名	張 雅婷（Zhang Yating）
論文題目	A Study on Object Search and Relationship Search from Text Archive Data （テキストアーカイブデータからのオブジェクト検索と関係検索に関する研究）		

（論文内容の要旨）

本論文は、長期間にわたり記録されたテキストアーカイブ情報から、オブジェクトや関係を検索する方法論について論じたものである。

大規模なテキスト情報が、多様なドメインでアーカイブとして蓄積されている。これらのテキストアーカイブデータは、様々な学術分野での研究に資するものであるが、アーカイブデータを研究に用いる上では、収集ドメイン、時間的ドメイン、空間的ドメインが多種多様で異なっているため、語（概念）の同定・検索や、因果関係などの語（概念）間の関係の同定・検索が困難であるという課題が存在する。例えば、時間的に異なるドメインで収集されたデータでは、語（概念）の意味が変化していたり、一方のドメインでの語彙と他方のドメインでの語彙が異なったりしているため、ある語（概念）に対応する他ドメインの語（概念）を見つけることすら容易ではない。本論文は、このようなテキストアーカイブの検索や分析に関わる困難な課題を克服することを目的として研究を行ったものである。

本論文は全6章から構成されている。その概要は以下の通りである。

第1章は序論であり、本論文の研究の背景、本論文の研究を行うに至った動機、ならびに本論文の研究の全体の概要を述べている。

第2章では、第3章から第5章において説明される研究課題に関して、それらに深く関連する分野の従来研究を整理し、本研究との位置付けを議論している。

第3章では、異なる複数のドメイン、すなわち、異なる語彙の意味空間を横断して、意味類似する語（概念）を検索するオブジェクト検索の問題について検討を行っている。例えば、iPodに意味類似する1980年代のデバイスを検索したい場合に、解候補となり得るWalkmanという語を検索ユーザが知らなければ、従来の検索システムでは検索すら実行出来ない。この問題は、起点領域（source domain）から目標領域（target domain）へのメタファー的写像をどのようにして求めるかという問題にも相当する。この問題に対して、本章では、異なる語彙の意味空間に対して、一方の意味空間における語（概念）に対応する他方の意味空間上の語（概念）を発見するために、一方の意味空間を変換する効率的な方法を提案している。

第4章では、テキストアーカイブ中での、語（概念）の意味的類似性を説明するのに根拠となる語（概念）集合を求める問題・手法について論じている。特に、語（概念）対の意味的類似性を支持する語対集合を自動的・効率的に求める手法を提案している。具体的には、入力語対と候補語対に対して語間の適合性、意味的類似性、関係類似性を計算する手法、及び、さらに候補語対集合における重要度を計算する手法を提案し、これらの手法の有効性を検証している。

第5章では、語（概念）の意味が時間的変化する状況下において、語（概念）の意味の時間的変化を抽出する手法、および、この時間的変化も考慮した上での、語（概念）の因果関係を抽出する手法を提案している。提案手法は、語（概念）の出現頻度・利用頻度の変動を分析し、これに基づき、ある語（概念）集合の出現が別の語（概念）集合の生起を引き起こしているかを検出するものである。

第6章では、本研究で得られた研究成果をまとめ、さらに今後の展開について議論している。

注）論文内容の要旨と論文審査の結果の要旨は１頁を３８字×３６行で作成し、合わせて、３，０００字を標準とすること。

論文内容の要旨を英語で記入する場合は、４００～１，１００wordsで作成し

審査結果の要旨は日本語５００～２,０００字程度で作成すること。
（続紙 ２ ）

（論文審査の結果の要旨）

本論文は、長期間にわたり記録されたテキストアーカイブ情報からオブジェクトや関係を検索する手法を提案し、大規模なテキストアーカイブデータを用いた評価実験を行い、提案手法の有効性を確認している。

長期間、多所で記録・蓄積されてきたテキストアーカイブにおいてオブジェクトや関係の検索をいかに効果的・効率的に行うかという問題は、その重要性にもかかわらず、情報検索分野においてすら従来研究が少ない研究課題である。

本論文は、テキストアーカイブ情報の検索に必要となる検索技術の同定・実現を目的として、特に、テキストアーカイブ情報からのオブジェクトの検索および関係の検索に焦点をあてて研究を行ったものである。テキストアーカイブ情報からの情報検索を困難にしている原因として、異なる時間や異なるドメインで記録・蓄積された情報であるために生じる「用語ギャップ（terminology gap）」の問題がある。用語ギャップ問題は、端的には、テキスト上に出現する語（概念）の意味が、時間やドメインによって異なるという問題である。このため、オブジェクト名による検索や、複数のオブジェクト間の関係を検索する際に、類似性判定を難しくしている。

本論文は、このために、テキストアーカイブからオブジェクトや関係の検索の性能を向上させるための検索方式を提案し、実際の大規模なテキストアーカイブ情報に適用し、その提案方式の有効性を検証している。

具体的には、本論文の研究によって得られた成果は以下のように要約される。

1. 異なるドメインで収集･記録された語彙集合間に存在する用語ギャップ問題を克服する目的で、語（概念）の意味空間を教師無しで変換する汎用的な手法を提案した。この変換技術により、２つの異なる時間期間に収集記録されたテキストアーカイブを対象として、一方のアーカイブに出現する語（概念）に対して、他方のアーカイブ内で意味的に類似する語（概念）を検索することが可能となった。提案手法は、具体的には、2つの異なるテキストアーカイブの語彙を、Skipgramモデルに基づく語の意味の分散表現技術を用いてそれぞれ学習し、２つの意味ベクトル空間を生成する。次に、両意味ベクトル空間上で、意味変化が比較的小さいと想定される語集合を検出し、このような語集合を用いて一方の意味ベクトル空間を変換する行列を求めるというものである。意味ベクトル空間の変換については、意味ベクトル空間の全体構造および部分構造を考慮した変換手法を提案している。長期間にわたって記録されたテキストアーカイブ（ニューヨークタイムズコーパス20年分、タイムズアーカイブ200年分）を対象として、提案の変換手法の評価実験を行い、提案手法の有効性を確認している。
2. 語の意味の分散表現法で学習して生成される語（概念）の意味ベクトル空間の妥当性を評価するための評価手法を提案した。これにより、テキストアーカイブ上での2つの語（概念）の間の意味的類似性を評価できるようになった。提案手法は、意味的類似性を判定したい語の対（入力語対）を与えると、この入力語対の意味的類似性を支持するような候補語対を求めるものである。入力語対と候補語対に対して、語間の適合性、意味的類似性、関係類似性を計算する手法（quality-based method）、及び、さらに候補語対集合内における候補語対の重要度の計算を追加した手法(systematicity-based method)の2つを提案し、語（概念）の対応度指数（top counterpart view）、および、類似性説明指数（similarity explanation view）という2種の評価

尺度を用いて、ニューヨークタイムズのアーカイブデータを対象として提案手法の評価実験を行い、提案手法の有効性を確認している。

3. 語（概念）の意味が時間的変化する状況下において、語（概念）間の因果関係を効率的に抽出する手法を提案している。提案手法によって抽出出来る因果関係は、従来のものとは異なり、語（概念）単位での因果関係抽出を行える、すなわち、token-levelの因果関係を発見する手法となっている。これにより、テキストアーカイブに内在する、語（概念）間の因果関係を精度良く抽出することが可能になった。さらに、本章で提案した因果関係抽出手法を、18年間にわたって記録されたアマゾン製品レビューデータに適用し評価実験を行い、実際に、興味深い因果関係の抽出に成功している。評価実験では、製品技術の進展が社会生活の質にどのような影響を及ぼしているかに着目し、これに関連する因果関係の抽出を試みている。このために、語彙を、製品技術に関する語彙と製品の利用に関する語彙の2種類に分類し、両語彙間の因果関係の抽出を試みている。

本論文は、長期間にわたり記録・蓄積されたテキストアーカイブから、類似のオブジェクトや関係を語（概念）単位で検索する有効な検索手法を提案しており、今後の情報検索に関する研究において重要な位置を占めると考えられる。このように、本論文は、学術上寄与するところが少なくないため、博士（情報学）の学位論文として価値あるものと認める。また、平成28年8月10日に論文内容とそれに関連した事項について試問を行った結果、合格と認めた。

注)論文審査の結果の要旨の結句には、学位論文の審査についての認定を明記すること。更に、試問の結果の要旨（例えば「平成　年　月　日論文内容とそれに関連した口頭試問を行った結果合格と認めた。」）を付け加えること。

Webでの即日公開を希望しない場合は、以下に公開可能とする日付を記入すること。

要旨公開可能日：　　年　　月　　日以降